Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 8846634, 13 pages
https://doi.org/10.1155/2021/8846634

Research Article

A Study on Travel Time Estimation of Diverging Traffic Stream on Highways Based on Timestamp Data

Sunghoon Kim,[1] **Hwapyeong Yu**,[2] **and Hwasoo Yeo**[2]

[1]*The Korea Transport Institute, Sejong, Republic of Korea*
[2]*Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea*

Correspondence should be addressed to Hwasoo Yeo; hwasoo@gmail.com

Received 24 March 2020; Revised 24 November 2020; Accepted 19 January 2021; Published 28 January 2021

Academic Editor: Rakesh Mishra

Travel time is valuable information for both drivers and traffic managers. While properly estimating the travel time of a single road section, an issue arises when multiple traffic streams exist. In highways, this usually occurs at the upstream of diverge bottleneck. The aim of this paper is to provide a new framework for travel time estimation of a diverging traffic stream using timestamp data only. While providing the framework, the main focus of this paper is on performing a few analyses on the stage of travel time data classification in the proposed framework. Three sequential steps with a few statistical approaches are provided in this stage: detection of data divergence, classification of divergent data, and outlier filtering. First, a divergence detection index (DDI) of data has been developed, and the analysis results show that this new index is useful in finding the threshold of determining data divergence. Second, three different methods are tested in terms of properly classifying the divergent data. It is found that our modified method based on the approach used by Korea Expressway Corporation shows superior performance. Third, a polynomial regression-based method is used for outlier filtering, and this shows reasonable performance even at a relatively low market penetration rate (MPR) of probe vehicles. Then, the overall performance of the travel time estimation framework is tested, and this test demonstrates that the proposed framework can show improved performance in distinctively estimating the travel times of two different traffic streams in the same road section.

1. Introduction

The growth in urban population and city-centred life patterns has raised a series of problems in modern cities such as traffic congestion and accidents. To tackle these issues, there have been numerous attempts to implement intelligent transportation system (ITS) in the road networks. In the field of ITS, travel time is one of the most valuable information for both vehicle drivers and traffic managers. In advanced traveler information system (ATIS), updating travel time for drivers in real time enables them to make informed decisions on their route choices to avoid congested roads [1]. In advanced traffic management system (ATMS), based on proper analyses of traffic states in relation to travel time information, traffic managers can develop various control and operational strategies to reduce road congestion [2]. Furthermore, travel time information can be used as supplementary input for further development of the recent studies related to traffic flow analysis [3], including short-term prediction using neural network techniques [4, 5] and long-term forecasts using fuzzy theory [6, 7]. Hence, proper estimation of travel time on each road in a real-time manner is crucial for further development and implementation of ITS.

Travel time information can be obtained indirectly or directly. The indirect ways usually estimate travel time based on traffic flow and speed measured by point sensors on the roadside such as loop detectors, video cameras, and radars [8–10]. The direct ways measure the travel time using records at tollgates and roadside units (RSU), which are parts of automatic vehicle identification (AVI) technologies [11]. The use of Global Positional System (GPS) on the probe vehicles equipped with navigation devices or smartphones is another direct way of travel time measurement, and this is the recent

technology increasingly used due to wide spatial coverages and low operational costs [12, 13].

There are three crucial issues in using the probe vehicles for travel time estimation. The first is the market penetration rate (MPR). The estimation is usually made by attempting a statistical analysis of the travel time distribution of traveling vehicles, but the estimation performance is reduced when the number of the probes is low. Hence, there were a few studies that examined the minimum conditions of MPR [14, 15]. There were also some studies that provided techniques for properly estimating travel time even with low MPR conditions [16–18]. The second issue is the outlying observations in collected probe data. The outliers can occur due to various reasons such as en route stops, measurement errors, and multiple devices in a single vehicle. Thus, there were several studies that attempted to solve this issue [19–22]. The third issue arises where different traffic streams exist on a single road section. In highway networks, this occurs at the upstream of a diverging road section, particularly when the disparity of the demands towards two traveling directions is large. This is called diverge bottleneck [23, 24], and an example is shown in Figure 1. In urban road networks, such cases usually occur at most intersections due to signal controls. Regardless of the road types, when such cases occur, there is a high risk of providing insufficient travel time information to drivers. For example, if the mean value of travel time were simply calculated from the data in Figure 2 and given to the drivers who are to travel through the off-ramp (left-turn), the traffic situation would be worse than if they were informed as they arrived at the site. To tackle this issue, separately estimating the travel times for each of the diverging streams is required.

In fact, there are several previous studies that made some efforts in relation to the third issue [16, 25–27]. Their focus was usually on estimating the vehicle delays at the diverging (or turning) spot on a road, and this was done separately from estimating the travel time of forward stream on the road before the diverging spot [16]. A limitation still exists, as they express the travel time with a single value by combining the two different properties. Such works may support the traffic control operations but may not be suited for improving the route guidance system. To make improvements in this issue, a few recent studies attempted to estimate the travel times of different stream directions distinctively [13, 28]. These works use GPS-based vehicle trajectories, and the advantage of them is that there is a higher chance of increasing estimation accuracy by microscopically tracking vehicle positions. However, the real-time feasibility may be an issue because they require continuously tracking down the traveling directions and trajectories of a number of vehicles based on numerous GPS points, which is a time-consuming and complex process. In the sense of real-time implementation, both the input data and the estimation technique should maintain simplicity. This is why many of the related studies stuck to the use of statistical analyses of travel time distribution based on the simplified timestamp data of inbound and outbound vehicles on the roads [15, 19, 29]. The major problem of using such simplified data is still that the estimation accuracy may be crucially influenced by the number of data points collected. Hence, each of the methods using trajectory data and timestamp data has both advantages and disadvantages depending on the environments of collecting and processing traffic data. Nonetheless, considering the motivation of this study, which is to improve ATIS, it is needed to maintain the simplicity of data processing while attempting to increase the estimation accuracy with insightful approaches.

Therefore, the aim of this paper is to provide a framework for travel time estimation of a diverging traffic stream using timestamp data only. While providing the framework, the main focus of this paper is on conducting a few insightful analyses on the stage of travel time data classification in the proposed framework. The analyses are done with the aid of a microsimulation program that allows us to test the estimation performance in various conditions of data acquisition. The scope of this paper deals with a diverging road section of highways, as the initial work of the estimation framework. When this initial work is successfully practiced, the research scope can be extended to dealing with the traffic at signalized intersections in urban road systems.

The rest of this paper is organized as follows. Section 2 provides the framework of travel time estimation of a diverging traffic stream. Section 3 describes the simulation settings and simulated data used for the analyses. Section 4 presents the results of analyses on the proposed estimation framework. Then, Section 5 concludes this paper and offers a few suggestions for further works.

2. Proposed Framework

Figure 3 shows the framework for travel time estimation of a diverging traffic stream. The timestamp data of inbound and outbound vehicles on the roads are used for the input data. Based on the collected input data, the travel time values of a road are sampled at each given time interval (e.g., every 5 minutes) in the preprocessing stage. Then, by conducting a few statistical analyses, the sampled travel time values are classified into forward and turning streams (or off-ramp stream). Next, the outputs are the travel time values of the forward stream and turning stream that are distinctively estimated based on the classified data.

In this paper, our focus is on doing a few analyses in terms of developing the stage of travel time data classification. Three sequential steps with a few simple statistical approaches are provided in this stage: detection of data divergence, classification of divergent data, and outlier filtering. The details of the steps are presented in the following sections.

2.1. Detection of Data Divergence. The purpose of this first step is to detect the divergence in travel time distribution and make a judgment if the divergence is significant. In the sense of conventional distribution analysis [19], there is a high risk of considering the sparse data points (red and green points) in Figure 2 as outliers, even though they should be considered as the effect of traffic state changes in the turning stream. Hence, it is necessary to provide an analysis to detect

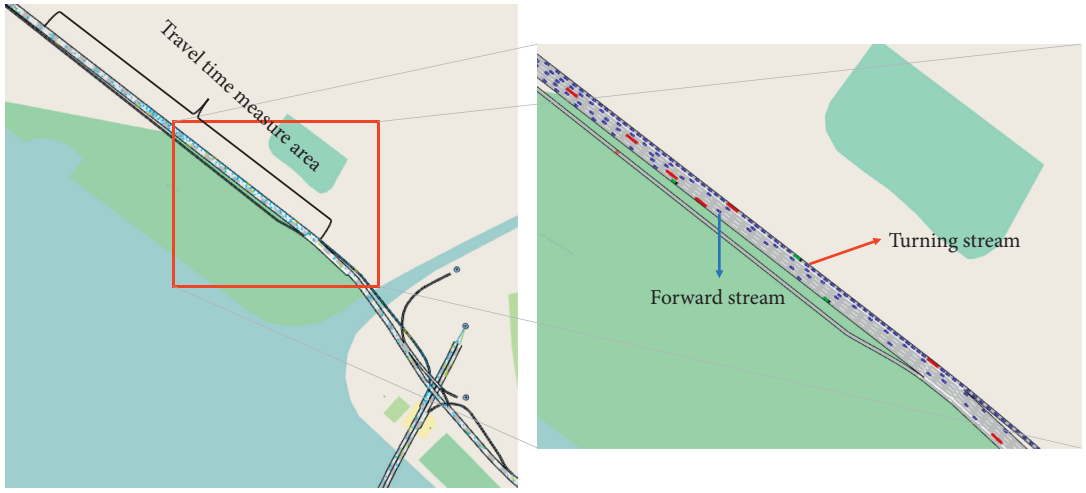

Figure 1: Example of diverge bottleneck on highways.

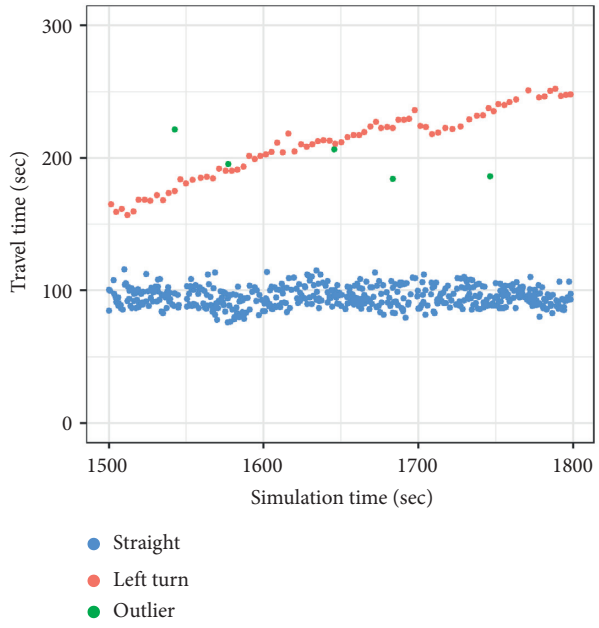

Figure 2: Travel time data of different traffic streams at the upstream of diverge bottleneck.

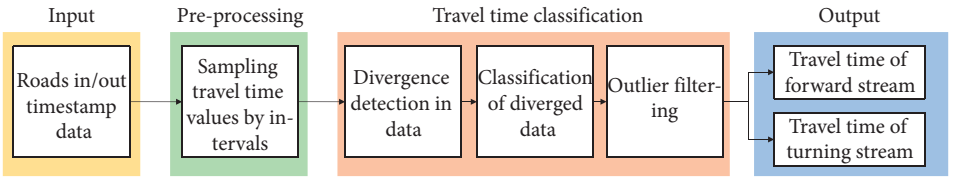

Figure 3: Framework for travel time estimation of diverging traffic stream.

whether the sparse data points are outliers or the congestion has an effect on the turning stream.

In general, it is known that the mean value and median value of distribution represent different levels of statistical significance. The arithmetic mean value represents the average behavior that is easily skewed even by a small proportion of extreme values. On the other hand, the median value is particularly useful in analyzing datasets with some proportion of extreme values because it is not much skewed by the extreme values. Hence, several previous studies used the basic advantage of comparing or combining these two properties for travel time estimation, considering the effect of extreme values in data [12, 21, 29].

If both the forward stream and turning stream are free flow or congested, the streams can be considered as a whole stream since they share the same traffic state. In this case, the frequency distribution diagram of travel time usually shows a unimodal shape with certain skewness [21]. Depending on the skewness, the difference in the mean and median values changes over time, but the amount of difference would not be significant over the entire distribution. On the other hand, if either one of the streams is free flow and the other is congested as in Figure 1, the distribution diagram of travel time would have a bimodal shape with two peaks as in Figure 4. The height of the first peak (the peak on the left side) with lower values would more likely be greater than that of the other peak, because the travel volume of free flow stream is higher. In this case, the mean value would be between the two peaks, and the median value would be located near the first peak. Hence, if the distribution shows clearly separated bimodal shape, the amount of difference in the mean and median values would be more significant than the former case.

Here, some may raise a question about the case when the off-ramp volume is much lower than the forward volume. In this case, the congestion in the main (forward) stream would most likely affect the turning traffic at the upstream of the diverging section as well, even though the off-ramp itself is in free flow state. Thus, at the upstream of the diverging section, both the forward and turning streams would be congested in this case, meaning that the case when the forward stream is congested and turning stream is free flow would seldom occur. In other words, the examples shown in Figures 1, 2, and 4 would be the general case of diverge bottleneck in highway sections that occurs particularly when the disparity of the demands towards two traveling directions is large.

Now, let us use the statistical fundamentals described above for detecting the divergence in travel time distribution. Table 1 provides descriptions of related variables. On any highway section h, let N be the number of sampled data points within the time interval t_{i-1} to t_i. In addition, let $T_h^n(t_i)$ be each of the travel time values in h within a time interval where $n = 1, 2, 3, \ldots, N$, and the value n is given in the order of observation time. The observation time of a vehicle is the time when the stamp of outbound time is recorded as the vehicle finishes its trip in h. Then, let us provide the index for detecting whether the divergence is significant and the sparse data points should not be filtered as outliers.

$$D_h(t_i) = \frac{\left|T_h^{me}(t_i) - T_h^{md}(t_i)\right|}{s_h(t_i)}, \quad (1)$$

where $T_h^{me}(t_i)$ is the arithmetic mean value, $T_h^{md}(t_i)$ is the median value, and $s_h(t_i)$ is the standard deviation value of $T_h^n(t_i)$. The index, $D_h(t_i)$, is the ratio of the difference between the mean and median values over the standard deviation, and we call this the divergence detection index (DDI) of data. The difference value is divided by the standard deviation in order to give weight to the significance of the difference. For the same difference value, if the data dispersion is large, the difference would be less significant out of the entire dataset. Conversely, if the data dispersion is small, the difference should be considered as more significant. In this way, we can make judgments on the significant level of the data divergence more properly. If DDI value is greater than a given threshold, the divergence in data is judged to be significant and the sparse data points should not be filtered as outliers, because the difference between the two properties over the dispersion of the distribution is considerably large. Then, the classification is required to distinctively deal with the different behaviors of the traffic streams in two directions.

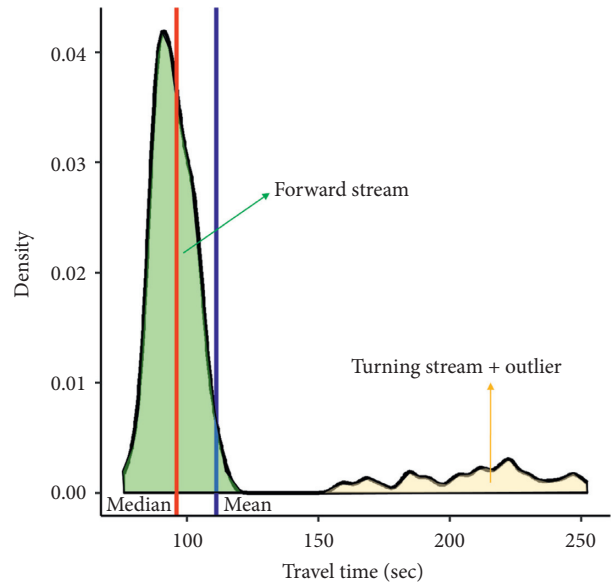

Figure 4: Bimodal shape of travel time data distribution.

Note that the threshold can differ for different road sections. Hence, an analysis is carried out for finding the threshold value of an example site in Section 4. The threshold may differ at different times of the day even at the same location, and examination on this should be done in further studies after this study is successfully practiced.

2.2. Classification of Divergent Data. When data divergence is detected by DDI and a given threshold, the data points are to be distinctively treated to estimate the travel times of the traffic streams in different directions. The step of classification of divergent data is provided for such a purpose. The major focus here is grouping the set of sampled travel time values into the forward stream and turning stream. At this step, the outliers would be still included in the group of the turning stream, and filtering them out is to be done in the following step. As can be seen in Figure 4, if we attempt to filter the outliers first, there is a high risk of treating the turning stream data points as outliers due to the low density of the data points. This would lower the accuracy of estimating the travel time of the turning stream. Hence, the proposed framework in this study attempts to initially separate the data points into two groups at this current step and then proceed to the following step of filtering outliers from the group of turning stream data points. Such an

Table 1: Descriptions of related variables.

Notation	Description
h	Highway road section for travel time estimation
t_i	Time for data sample (at i-th interval)
N	Number of sampled data points within the time interval t_{i-1} to t_i
n	Order of data observation ($n = 1, 2, 3, \ldots, N$)
$T_h^n(t_i)$	n-th observed travel time value in h within a time interval
$T_h^{me}(t_i)$	Arithmetic mean of travel time values observed within a time interval
$T_h^{md}(t_i)$	Median of travel time values observed within a time interval
$s_h(t_i)$	Standard deviation of travel time values observed within a time interval

approach is the major difference in the proposed framework from the existing methods.

In this section, let $G_h^1(t_i)$ be the set of travel time values grouped as the turning stream (and outliers) and $G_h^2(t_i)$ be the set of travel time values grouped as the forward stream. Then, we classify $T_h^n(t_i)$ values into these two groups based on a given classification rule. The issue here is that there is no clearly superior solution, to the best of the authors' knowledge. The classification performance can differ depending on many factors such as what kind of statistical approaches is used, type of road, time of day, and sampled rate of data. In this paper, three methods are attempted to investigate the classification performances and find a reasonable solution. Note that all these methods are designed for the general case of diverge bottleneck in highway road sections, which has been described in Section 1.

2.2.1. Method 1. This method is designed to see the direct effects of the mean and median values for classification. In this method, we calculate both $T_h^{me}(t_i)$ and $T_h^{md}(t_i)$ first. Then, $T_h^n(t_i)$ is classified as turning stream (and outliers) if its value is closer to $T_h^{me}(t_i)$. Conversely, $T_h^n(t_i)$ is classified as forward stream if its value is closer to $T_h^{md}(t_i)$ as follows:

$$G_h^1(t_i) = \left\{T_h^n(t_i):\ \left|T_h^n(t_i) - T_h^{me}(t_i)\right| < \left|T_h^n(t_i) - T_h^{md}(t_i)\right|\right\},$$
$$G_h^2(t_i) = \left\{T_h^n(t_i):\ \left|T_h^n(t_i) - T_h^{me}(t_i)\right| \geq \left|T_h^n(t_i) - T_h^{md}(t_i)\right|\right\}. \quad (2)$$

2.2.2. Method 2. In this method, we simply use the mean value $T_h^{me}(t_i)$ and the standard deviation $s_h(t_i)$. We assume that the data points beyond a certain range of the mean value are considered as the data originated from different behaviors. In this method, we consider that the range is 1.5 times the standard deviation. $T_h^n(t_i)$ is classified as the turning stream (and outliers) if its value is beyond the given range, and it is classified as the forward stream if its value stays within the range as follows:

$$G_h^1(t_i) = \{T_h^n(t_i):\ T_h^n(t_i) > T_h^{me}(t_i) + 1.5 \cdot s_h(t_i)\},$$
$$G_h^2(t_i) = \{T_h^n(t_i):\ T_h^n(t_i) \leq T_h^{me}(t_i) + 1.5 \cdot s_h(t_i)\}. \quad (3)$$

According to the "three-sigma rule," if the specific range is set to twice the standard deviation, about 95% of the data points would be included in the range. Then, the classification of the data points corresponding to the turning stream is more likely to be neglected, and this would lead to underestimation. Conversely, if the specific range is set to one times the standard deviation, about 68% of the data points would be included in the range, resulting in a highly exclusive classification. Then, a number of the data points corresponding to the forward stream are more likely to be incorrectly included in the group of turning stream data points, and this would lead to overestimation. Hence, the specific range should be chosen between 1 and 2 for avoiding both underestimation and overestimation, and as the base case study, the specific range is set to 1.5 times the standard deviation in this paper. Note that alternative values can be chosen depending on the condition of sampled data, and properly determining the specific range is worthy of further investigation in subsequent studies.

2.2.3. Method 3. The third method is based on the Korea Expressway Corporation (KEC), which is currently used for filtering outliers in probe data collected from the entire Korean highway system [19]. The reliability of the travel time estimated with this has been shown by the ATIS in South Korea almost for a decade, with empirical history. A coefficient of variation (CV) is determined as the ratio of the standard deviation $s_h(t_i)$ over the mean value $T_h^{me}(t_i)$. Then, the methods for removing outliers are provided for each CV condition as in Table 2.

As can be seen in the table, the original method removes the observed data points that are outside the predefined range, and the predefined range differs according to the CV condition. With respect to the same mean value, if CV is low, this means that the data dispersion is low; thus, only a small range of the data points are considered to be removed. As CV gets greater, the range of removing the data points increases due to higher data dispersion. However, in this paper, we consider the divergent data to be classified as the turning stream, rather than filtering them as outliers. Hence, in this paper, the data to be removed based on the conditions of the table are instead grouped as turning stream (and outliers) $G_h^1(t_i)$, and the other data values are grouped as the forward stream $G_h^2(t_i)$.

2.3. Outlier Filtering. As in Figure 2, not all travel time values in $G_h^1(t_i)$ are an actual part of the turning stream. The outliers may still exist within this group due to various reasons such as en route stops, measurement errors, and

Table 2: Outlier removal method of Korea Expressway Corporation.

Conditions	Removal method
CV < 0.05	Remove the top 2% and bottom 3%
0.05 ≤ CV < 0.10	Remove the top 5% and bottom 5%
0.10 ≤ CV < 0.15	Remove the top 8% and bottom 7%
CV ≥ 0.15	Remove the values outside mean ± standard deviation

multiple devices in a single vehicle. The purpose of this third step is to remove these outliers in order to increase the performance of estimating the travel time of the turning stream.

For filtering outliers from $G_h^1(t_i)$, an approach based on polynomial regression is taken in this paper. As can be seen in Figure 2, the divergent data points have a certain growth trend different from the data of the main stream, and this is due to the propagation of congestion shockwave in the upstream lane of the off-ramp. After the growth, the trend of divergent data points decreases as the congestion shockwave begins to dissipate. Hence, the rate of shockwave in the upstream lane of the off-ramp affects the travel time growth in time-series. Analyzing the trend of such a response variable (travel time) in time-series can be considered as a regression problem [30, 31].

Let $T_h^m(t_i)$ be each of the travel time values in $G_h^1(t_i)$ where $m = 1, 2, 3, \ldots, M$. Also, let $f_h^m(t_i)$ be the expected value of $T_h^m(t_i)$ in terms of the value of t, and t is in the unit of seconds within the range of t_i. Then, the quadratic equation of fitting the data points in $G_h^1(t_i)$ can be modeled as follows:

$$f_h^m(t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2, \quad \forall t \in t_i. \tag{4}$$

Let $G_h^3(t_i)$ be the data points to be grouped as outliers and $G_h^4(t_i)$ be the data points to be grouped as the turning stream, which makes $G_h^1(t_i) = G_h^3(t_i) \cup G_h^4(t_i)$. Here, let r_h be the constant factor that decides the range of outlier filtering based on the trend line $f_h^m(t_i)$. Then, $T_h^m(t_i)$ is classified as outliers if its value is beyond the range of r_h away from the trend, and it is classified as the turning stream if the value stays within the range as follows:

$$\begin{aligned} G_h^3(t_i) &= \{T_h^m(t_i) : T_h^m(t_i) > f_h^m(t) + r_h\} \cup \{T_h^m(t_i) : T_h^m(t_i) < f_h^m(t) - r_h\}, \\ G_h^4(t_i) &= \{T_h^m(t_i) : T_h^m(t_i) \leq f_h^m(t) + r_h\} \cap \{T_h^m(t_i) : T_h^m(t_i) \geq f_h^m(t) - r_h\}. \end{aligned} \tag{5}$$

Note that the range value r_h can differ for different conditions such as site location and traffic state. Hence, an analysis is done for finding the appropriate values for these properties in Section 4.

3. Data for Analyses

To analyze the performance of the proposed framework, it is necessary to obtain the true values containing the turning manoeuvres of vehicles towards a diverging direction. The true values are to be compared with estimated values for the performance analyses. However, the real-world data obtained by probe vehicles do not represent the true values. Hence, we obtain travel time values with the aid of AIMSUN microsimulation [32]. In this study, the data of all simulated vehicles are considered as true values.

Figure 1 presents the simulated road geometry of the diverging section, which connects Gangbyeonbuk-ro and Seongsan bridge in Seoul, South Korea. Gangbyeonbuk-ro is a riverside highway, and Seongsan bridge connects the north and south of Han River. There is a high traffic demand from the north to the south of the Han River during peak hours. The congestion frequently occurs in the off-ramp section, and the diverge bottleneck affects the upstream of Gangbyeonbuk-ro. Hence, the road section that corresponds to the upstream of the diverging highway section is selected for the analysis of this study. To simulate this, we set the "travel time measure area" just before the off-ramp section, which has a length of 1 km.

The traffic demand for the simulation is set to be similar to the peak hour demand in the real world. The overall traffic demand increases until the starting time of the peak hour, and the maximum traffic demand is maintained during the peak hour. After peak hour, the overall traffic demand decreases, and the simulation is terminated when the congestion at the diverging section disappears. Table 3 presents the traffic demand of each stream direction over simulation time, and the number of vehicles that are to travel through the off-ramp is 1/3 of the overall traffic demand. To construct realistic traffic conditions, we set the truck and bus demands, and the proportion of these demands is 5% each. These types of vehicles have a longer length and lower desired speed. Furthermore, to test the performance of the outlier filtering step, some outliers are added to the simulation timestamp data. These outliers are generated by specific vehicles that travel through the park next to the study site, and the type of these outliers is considered as the en route stops [20].

When the simulation is ended, the section travel time and exit time of each vehicle are stored in the simulation output database. We extract the vehicles that pass through the area of travel time measure and classify the vehicles based on the section exit time by 5-minute unit. The 5-minute timestamp data are constructed by the section travel

Table 3: Traffic demand of each stream direction over simulation time.

Simulation time (hh : mm)	00 : 00–00 : 30	00 : 30–01 : 00	01 : 00–01 : 30	01 : 30–02 : 00
Forward stream volume (veh/h)	10850	10850	8920	4960
Turning stream volume (veh/h)	5425	5425	4460	2480

time of the classified vehicles, which are plotted in the order of increasing exit time. Through the random sampling of vehicles in 5-minute timestamp data, we analyze the performance of the proposed framework in various MPR conditions of probe vehicles.

To ensure the generality of the proposed framework, we make 30 repeated simulations while changing the generation seed of the simulation. The generation seed determines the generation time and route of each vehicle within the predetermined road geometry and traffic demand setting. Due to changes in the generation seed, the characteristics of the diverge bottleneck, such as travel time difference of each stream direction or duration of the phenomenon, can vary for each simulation. Total number of travel time data points collected through the 30 simulations is 357811, of which 315038 are forward stream data points and 42773 are turning stream data points. The average travel time of forward and turning streams is 102.01 and 209.61 seconds, respectively. Also, the standard deviation values are 29.26 and 136.73 seconds, respectively.

4. Results of Analyses

The analyses on each of the three steps in data classification are done according to the confusion matrix in Table 4. It is used for testing the performance of the methods with the number of correct and incorrect data groupings. The performance can be classified into four categories: true positive (TP), false positive (FP), false negative (FN), and true negative (TN).

Note that the meanings of the four categories vary in each of the analyses. The performance tests in each analysis will be done based on the true positive rates and true negative rates at different MPR levels. We assume that the entire set of the simulated travel time data described in Section 3 is the data obtained at 100% MPR, and they are considered as the actual results (true values). Then, we randomly sample the data points from the entire dataset for the lower MPR cases. Using the sampled data points at a given MPR condition, the true positive rate and true negative rate can be calculated as follows:

$$\text{True positive rate} = \frac{\text{TP}}{\text{TP} + \text{TN}},$$
$$\text{True negative rate} = \frac{\text{TN}}{\text{FP} + \text{TN}}. \quad (6)$$

4.1. Analysis of Divergence Detection. The first step is to detect the divergence of the traffic stream. The value of DDI represents the significant level of traffic divergence; thus, it varies depending on the traffic state at a diverging road section. The traffic state changes over the simulation time under the given travel demands in Table 3, and the traffic state at the same simulation time can have differences by the 30 simulations due to the generation seed that is given differently to each of the simulations. Figure 5 shows the boxplots of DDI values of 30 different simulations at every 5-minute sample interval. The red dashed lines show the time range when the divergence in travel time data exists. The boxplots show clear changes in the distribution of D_h values. The value is higher than 0.3 during the time range of data divergence for this specific road section. Hence, in this study, 0.3 is used as the threshold of detection of data divergence. The sampled travel time values at t_i are required to be classified, if $D_h(t_i) > 0.3$.

Table 4: Confusion matrix.

		Actual result	
		Positive	Negative
Estimation	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Figures 6(a) and 6(b) show the TP and TN rates at different MPRs when DDI threshold is 0.3, respectively. Here, the TP is the case when there is a divergence in actual data and the distribution of the sampled travel time values has $D_h(t_i) > 0.3$. TN is the case when there is no divergence in actual data and the distribution of the sampled values has $D_h(t_i) < 0.3$. As can be seen in these subfigures, both the TP and TN rates are less than 0.8 for the lowest MPR case. However, both rates increase as the MPR increases, and they show sufficient performances, which are above 0.8 of the rates, for MPR higher than 10%. These results show that, even at relatively low MPR conditions, the proposed method using the DDI can perform well for detecting divergence in travel time data.

4.2. Analysis of Data Classification. Figure 7 shows the results of the three classification methods described in Section 2. Here, TP is the case when the sampled data points are grouped into G_h^1 and these data points are the subset of the actual turning stream (including outliers). TN is the case when the sampled data points are grouped into G_h^2 and these data points are the subset of the actual forward stream.

In Figure 7(a), TP rates of all three methods increase as the MPR increases. They all show similar patterns and perform well even at relatively low MPR, because the TP rates are higher than 0.8 when MPR is greater than 15%. Out of the three methods, method 1 outperforms the others and method 2 shows the worst performance. On the other hand, in Figure 7(b), method 1 shows the worst performance in TN rates, while the other two methods show high performances even at low MPR conditions.

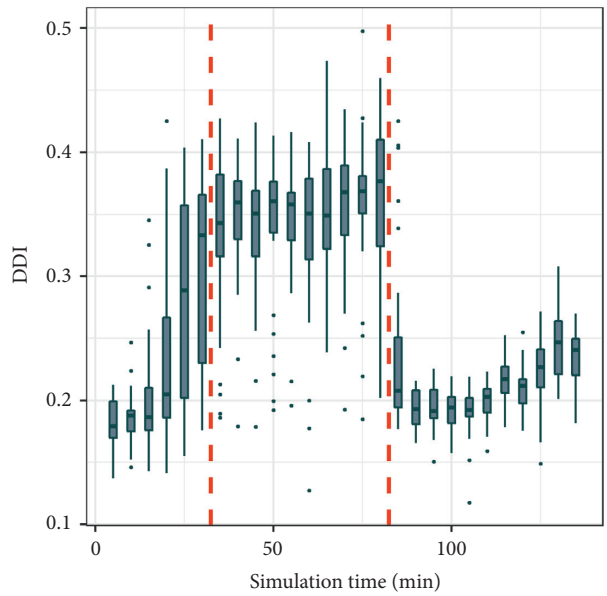

Figure 5: Boxplots of DDI values of 30 different simulations at every sampling interval.

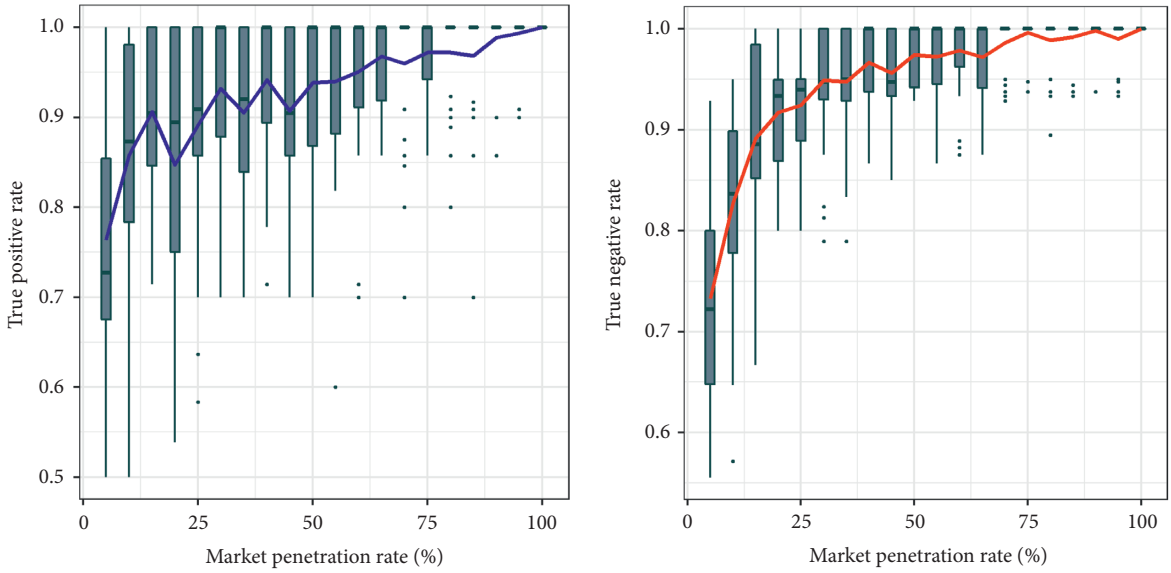

Figure 6: Results of data divergence detection (0.3 of DDI threshold). (a) TP rates at different MPRs. (b) TN rates at different MPRs.

This is because method 1 is the only one not considering the standard deviation of the distribution. Hence, it is not recommended that method 1 is used, based on the performances of both TP and TN. This finding definitely suggests the importance of considering the standard deviation of statistical distributions in the field of data classification. Now, if we check the other two, method 3 outperforms method 2 in TP rates, and it shows lower performance than method 2 in TN rates. However, the difference in TN rates is very slight. Hence, method 3 is superior based on both TP and TN rates. Thus, the analyses in the following sections are done based on the results of data classification derived by method 3.

4.3. Analysis of Outlier Filtering. This section presents the results of the outlier filtering analysis. Here, TP is the case when the sampled data points are grouped into G_h^3 and these data points are the subset of the actual outliers. TN is the case when the sampled data points are grouped into G_h^4 and these data points are the subset of the actual turning stream.

While applying the regression model $f_h^m(t)$, notice that, for example, the values of regression coefficients are $\beta_0 = 2.846 \times 10^2$, $\beta_1 = 8.368 \times 10^{-1}$, and $\beta_2 = -2.871 \times 10^{-3}$ when MPR is 100% for the 11th data collection interval of a simulation case, and these values vary for all different analyses conditions (30 different simulations × 24 data collection intervals × 20 MPR conditions).

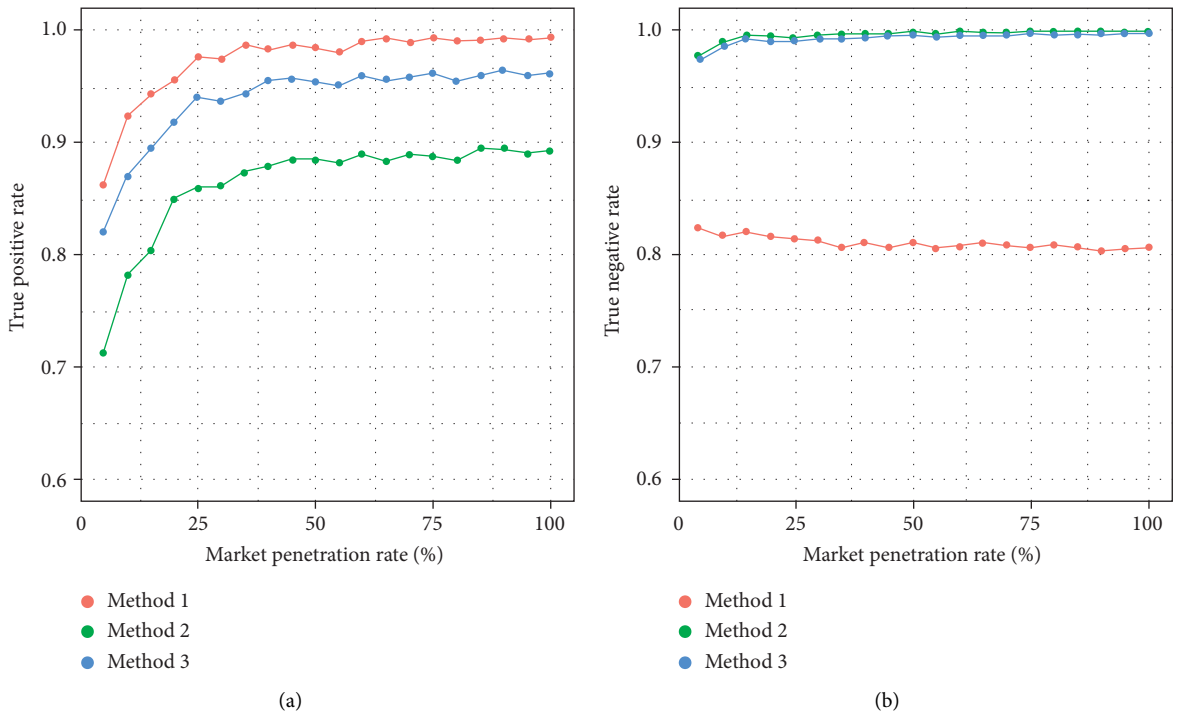

Figure 7: Results of the three classification methods. (a) TP rates at different MPRs. (b) TN rates at different MPRs.

Figure 8 shows the TP and TN rates when different values are applied to r_h. This graph shows that the TP performance is superior as the factor r_h decreases. This is because, for lower r_h, there is a higher chance of including most of the actual outliers into G_h^3. However, this would include the actual turning stream data points into the same group and may lead to a high error. On the other hand, the TN rates show that r_h should be at the greatest level for higher TN performances. This is because, for higher r_h, there is a higher chance of including most of the actual turning stream data points into G_h^4. However, this would include the actual outliers into the same group and may lead to a high error. Hence, by considering both TP and TN rates in this graph, it is conclude that the appropriate r_h value is 27 for balancing TP with TN performances and maintaining both at a high level.

Figure 9 shows the TP rates and TN rates at different MPRs when $r_h = 27$. The TP rates in Figure 9(a) show an increasing trend as MPR gets greater, and they show reasonable performances, which are near 0.8, only when MPR is higher than 20%. The low performances at low MPRs are because some of the actual outliers are located within the range of the turning stream cluster as in Figure 2, and these points could not be filtered from the cluster. The TN rates in Figure 9(b) show reasonable performances near 0.8 for all MPR conditions. The performances are almost constant for all conditions except for 5% of MPR. The performance is exceptionally a bit higher for 5% of MPR, and this is because the number of actual outliers is extremely low within the sampled data points under the given MPR condition. Overall, the results imply that the outlier filtering shows reasonable performances even for relatively low MPR conditions.

4.4. Overall Results of Travel Time Estimation. The overall performance of travel time estimation is tested in this section. Figure 10 shows the parity plot of travel time estimation before and after applying the proposed framework. The x-axis is the actual travel time of the turning stream, and the y-axis is the estimated travel of the turning stream. The black solid line is the reference where the actual and estimated values match each other. In this parity plot, the estimation accuracy is higher as the data points are closer to the reference line. This figure presents the results of travel time estimation before and after applying the proposed framework for different actual travel time values from 70 to 340 seconds. As the travel time of the turning stream increases, the distance between the data points and the reference line gradually increases when the conventional method is used. Conversely, when the proposed estimation framework is applied, the distance between the data points and the reference line is significantly reduced. The root mean squared error (RMSE) values with respect to the reference line are decreased from 113.11 to 20.42 seconds. This shows clear evidence that the proposed framework has a high chance for accuracy improvement.

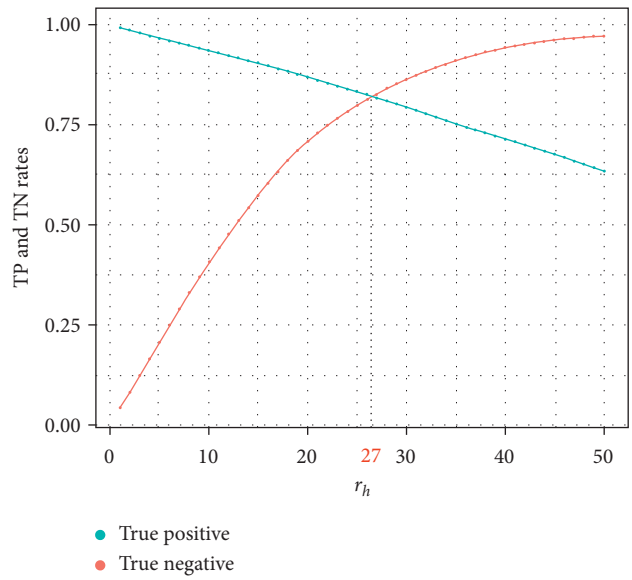

Figure 8: TP and TN rates at different r_h values.

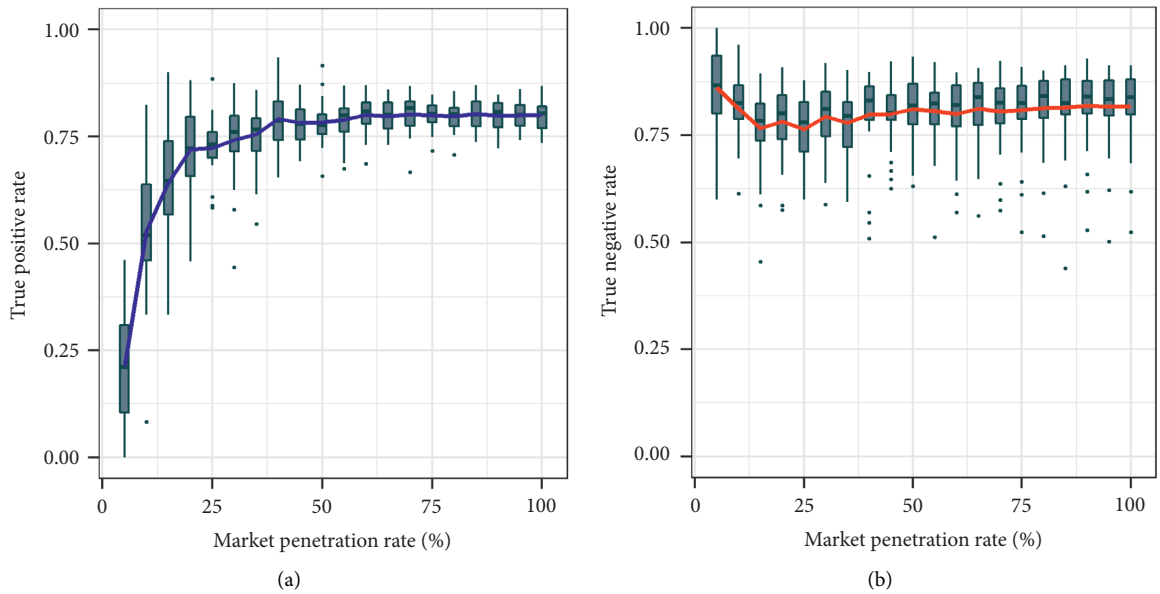

Figure 9: Results of outlier filter analysis. (a) TP rates at different MPRs when $r_h = 27$. (b) TN rates at different MPRs when $r_h = 27$.

Figure 10 depicts only a comparison result of a specific case (at MPR 30%) as an example, and the evaluation at various MPR conditions is also required. The evaluation is done by comparing the mean absolute percentage error (MAPE) values before and after applying the proposed framework with the three data classification steps. The MAPE values are calculated as follows.

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\text{actual}_t - \text{estimated}_t}{\text{actual}_t}\right| \times 100\,(\%). \quad (7)$$

We consider the mean value of $T_h^n(t_i)$ as the estimated travel time values of both forward and turning streams before applying the framework. For the "after" case, we carry out the output stage shown in Figure 3. The arithmetic mean value of the data points in $G_h^2(t_i)$ is the estimated travel time of the forward stream, and the arithmetic mean value of the data points in $G_h^4(t_i)$ is the estimated travel time of the turning stream.

Figure 11 shows the overall performance tests on estimating travel times of traffic streams in different directions.

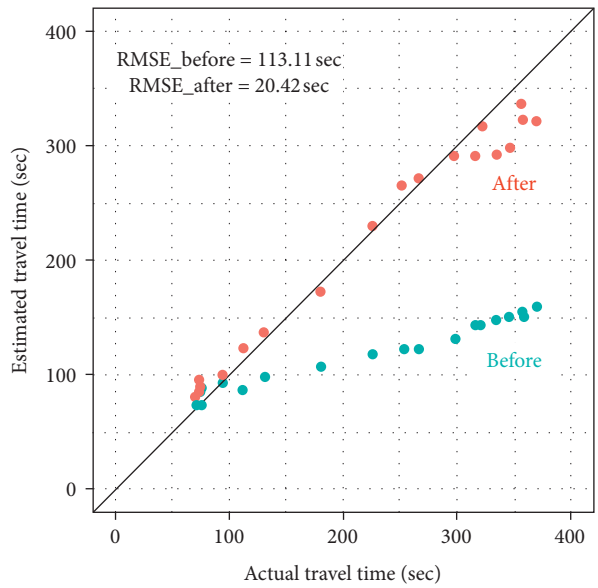

Figure 10: Parity plot of travel time estimation before and after applying the proposed framework (at MPR 30%).

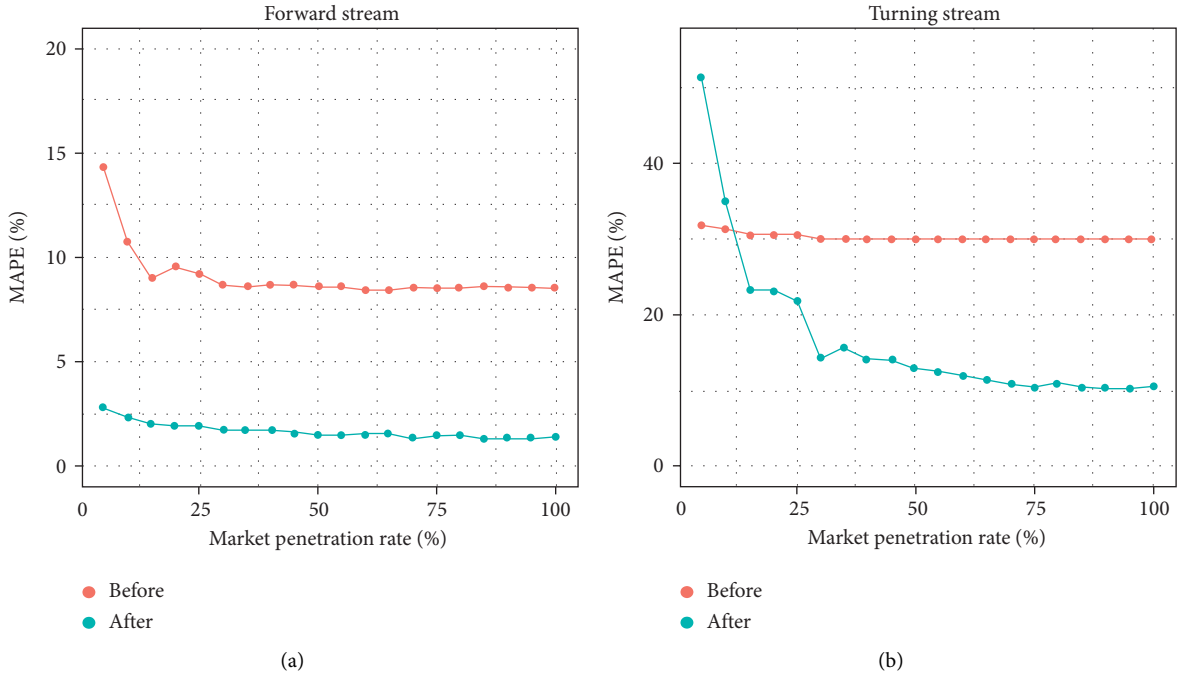

Figure 11: MAPE values of travel time estimation before and after applying the proposed framework. (a) Forward stream estimation at different MPRs. (b) Turning stream estimation at different MPRs.

The figure presents the MAPE values of travel time estimation before and after applying the proposed framework at different MPRs. For both streams, it is shown that the MAPE values are significantly reduced when the proposed estimation framework is applied, except for the turning stream cases with very low MPR conditions (5~10%). Such exceptions occur due to the low outlier filtering performance at low MPR conditions, and the reason for this was described in Section 4.3. Hence, the framework of distinctively estimating the travel times proposed in this study is considered to be suitable for use when MPR is more than 10%. If some improvements of the method for removing outliers are made in further studies, the entire framework proposed in this study is expected to show reasonable performance under the overall MPR condition. Subsequent studies shall particularly consider how to precisely filter the outliers located within the range of the turning stream data cluster for very low MPR conditions.

Still, for MPR higher than 10%, the MAPE of the turning stream in Figure 11(b) gradually decreases as the MPR increases. These results imply that the proposed estimation

framework can show improved performance in distinctively estimating the travel times of two different traffic streams in the same road section even for relatively low MPR conditions.

5. Conclusion

In this paper, a new framework for travel time estimation of a diverging traffic stream is provided. In this framework, the timestamp data of inbound and outbound vehicles on the roads are used for the input data. These data are sampled at each given time interval in the preprocessing stage. Then, by applying simple statistical analyses, the sampled travel time values are classified into forward and turning streams (or off-ramp stream). Next, the travel time values of the forward stream and turning stream in a single road section are estimated distinctively based on the classified data.

While providing the framework, the main contribution of this paper is that it offers a new approach in travel time data classification and carries out a few insightful analyses to test the performance of the approach. Three sequential steps with a few statistical approaches are provided in the stage of travel time data classification: detection of data divergence, classification of divergent data, and outlier filtering. First, a divergence detection index (DDI) of data is newly provided, and the analysis results show that the index can be useful in finding the threshold of determining data divergence. Second, after detecting data divergence, three different methods are tested in terms of properly classifying the divergent data. Method 3, which is modified based on a method currently used by Korea Expressway Corporation [19], proves to be superior and it shows sufficient performance in classifying the divergent data points. Third, a polynomial regression-based method is used for outlier filtering, and it shows reasonable performance even at a relatively low market penetration rate (MPR) of probe vehicles. Then, the overall performance of the travel time estimation framework is tested. This test demonstrated that the proposed estimation framework can show improved performance in distinctively estimating the travel times of two different traffic streams in the same road section.

There are two research values in the estimation framework proposed in this study. The first is that this framework seeks real-time practicality by using simplified data such as timestamp data. The second is that it suggests a new sequential composition of the three-step analysis to improve the accuracy of distinctive estimation of the travel times of diverging traffic streams. In the existing ATIS, the average value of travel time data points collected in road sections is usually calculated for information services. However, this has a limitation in displaying detailed information for each vehicle's traveling direction at a diverging road section. Hence, the demands of navigation service providers are increasing for new technological developments to improve the existing services. Furthermore, in terms of future autonomous vehicle operation, there are increasing demands for the development of various data collection and analysis solutions to implement a high-definition map (HD map) that requires traffic condition information for each lane of the road. The framework presented in this study is expected to be used for the advancement of ATIS and ATMS in the future, as it is a method that can contribute to the improvement of the route guidance services and to the realization of the HD map service for the operation of autonomous vehicles.

Even though the proposed framework has shown reasonable performance, this paper is not without limitations. This work focuses on analyzing only a few classification methods for the divergent data. Hence, some alternatives should be tested in further work for improving classification performance. Furthermore, this work shows a limitation in filtering outliers at very low MPR conditions. Hence, it is suggested that further efforts are made for solving this issue as well. Moreover, the scope of the estimation framework deals only with the diverging road section of highways. There is a need to extend the research scope to deal with the traffic at signalized intersections in the urban road system.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1A2B2002329).

References

[1] L. Yang and X. Zhou, "Optimizing on-time arrival probability and percentile travel time for elementary path finding in time-dependent transportation networks: linear mixed integer programming reformulations," *Transportation Research Part B: Methodological*, vol. 96, pp. 68–91, 2017.

[2] S. Kim, S. Tak, D. Lee, and H. Yeo, "Distributed model predictive approach for large-scale road network perimeter control," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 5, pp. 515–527, 2019.

[3] J. Zeng, S. Yu, Y. Qian, and X. Feng, "Expressway traffic flow model study based on different traffic rules," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 6, pp. 1099–1103, 2018.

[4] S. Guo, Y. Lin, S. Li, Z. Chen, and H. Wan, "Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting," *Ieee Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3913–3926, 2019.

[5] L. Liu, J. Zhen, G. Li et al., "Dynamic spatial-temporal representation learning for traffic flow prediction," *IEEE*

Transactions on Intelligent Transportation Systems, In press, 2020.

[6] R. M. Li, Y. F. Huang, and J. Wang, "Long-term traffic volume prediction based on K-means Gaussian interval type-2 fuzzy sets," *IEEE-CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1344–1351, 2019.

[7] R. M. Li, C. Y. Jiang, F. H. Zhu, and X. L. Chen, "Traffic flow data forecasting based on interval type-2 fuzzy sets theory," *IEEE-CAA Journal of Automatica Sinica*, vol. 3, no. 2, pp. 141–148, 2016.

[8] Q. Gan, G. Gomes, and A. Bayen, "Estimation of performance metrics at signalized intersections using loop detector data and probe travel times," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 2939–2949, 2017.

[9] C. Lu and J. Dong, "Estimating freeway travel time and its reliability using radar sensor data," *Transportmetrica B: Transport Dynamics*, vol. 6, no. 2, pp. 97–114, 2018.

[10] J. J. Tang, Y. J. Zou, J. Ash et al., "Travel time estimation using freeway point detector data based on evolving fuzzy neural inference system," *PLos One*, vol. 11, no. 2, 2016.

[11] M. L. Tam and W. H. K. Lam, "Using automatic vehicle idenification data for travel time estimation in Hong Kong," *Transportmetrica*, vol. 4, no. 3, pp. 179–194, 2008.

[12] J. Cheng, G. Li, and X. H. Chen, "Developing a travel time estimation method of freeway based on floating car using random forests," *Journal of Advanced Transportation*, vol. 2019, Article ID 8582761, 13 pages, 2019.

[13] C. Y. Shi, B. Y. Chen, and Q. Q. Li, "Estimation of travel time distributions in urban road networks using low-frequency floating car data," *Isprs International Journal of Geo-Information*, vol. 6, no. 8, 2017.

[14] A. Bolbol, T. Cheng, I. Tsapakis, and A. Chow, "Sample size calculation for studying transportation modes from GPS data," *Procedia—Social and Behavioral Sciences*, vol. 48, pp. 3040–3050, 2012.

[15] M. Cetin, G. F. List, and Y. J. Zhou, "Factors affecting minimum number of probes required for reliable estimation of travel time," *Data Initiatives*, vol. 1917, pp. 37–44, 2005.

[16] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transportation Research Part B: Methodological*, vol. 53, pp. 64–81, 2013.

[17] L. Tang, Z. Kan, X. Zhang, X. Yang, F. Huang, and Q. Li, "Travel time estimation at intersections based on low-frequency spatial-temporal GPS trajectory big data," *Cartography and Geographic Information Science*, vol. 43, no. 5, pp. 417–426, 2016.

[18] Y. Wang, Y. Zheng, and Y. Xue, "Travel time estimation of a path using sparse trajectories," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, August, 2014.

[19] J. Jang, "Outlier filtering algorithm for travel time estimation using dedicated short-range communications probes on rural highways," *Iet Intelligent Transport Systems*, vol. 10, no. 6, pp. 453–460, 2016.

[20] S. S. Moghaddam and B. Hellinga, "Evaluating the performance of algorithms for the detection of travel time outliers," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2338, no. 1, pp. 67–77, 2013.

[21] H. Park and Y. Kim, "Model for filtering the outliers in DSRC travel time data on interrupted traffic flow sections," *KSCE Journal of Civil Engineering*, vol. 22, no. 9, pp. 3607–3619, 2018.

[22] S. A. Shaikh and H. Kitagawa, "Continuous outlier detection on uncertain data streams," in *Proceedings ot the 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (IEEE Issnip 2014)*, Singapore, 2014.

[23] J. C. Muñoz and C. F. Daganzo, "The bottleneck mechanism of a freeway diverge," *Transportation Research Part A: Policy and Practice*, vol. 36, no. 6, pp. 483–505, 2002.

[24] J. Rudjanakanoknad, "Capacity change mechanism of a diverge bottleneck," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2278, no. 1, pp. 21–30, 2012.

[25] X. Ban, R. Herring, P. Hao, and A. M. Bayen, "Delay pattern estimation for signalized intersections using sampled travel times," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2130, no. 1, pp. 109–119, 2009.

[26] X. Liu, F. Lu, H. Zhang, and P. Qiu, "Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network," *Frontiers of Earth Science*, vol. 7, no. 2, pp. 206–216, 2013.

[27] I. Shatnawi, P. Yi, and I. Khliefat, "Automated intersection delay estimation using the input-output principle and turning movement data," *International Journal of Transportation Science and Technology*, vol. 7, no. 2, pp. 137–150, 2018.

[28] D. Wang, F. Fu, X. Luo, S. Jin, and D. Ma, "Travel time estimation method for urban road based on traffic stream directions," *Transportmetrica A: Transport Science*, vol. 12, no. 6, pp. 479–503, 2016.

[29] X. J. Ban, Y. Li, A. Skabardonis, and J. D. Margulici, "Performance evaluation of travel-time estimation methods for real-time traffic applications," *Journal of Intelligent Transportation Systems*, vol. 14, no. 2, pp. 54–67, 2010.

[30] J. Kwon, B. Coifman, and P. Bickel, "Day-to-day travel-time trends and travel-time prediction from loop-detector data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1717, no. 1, pp. 120–129, 2000.

[31] H. Y. Sun, H. X. Liu, H. Xiao, R. R. He, and B. Ran, "Use of local linear regression model for short-term traffic forecasting," *Initiatives in Information Technology and Geospatial Science for Transportation*, vol. 1836, pp. 143–150, 2003.

[32] AIMSUN, *AIMSUN Version 7 Users Manual*, TTS-Transport Simulation Systems, Barcelona, Spain, 2011.